Journal of Cellular Biochemistry 91:1197–1203 (2004)

Predicting Subcellular Localization of Proteins by Hybridizing Functional Domain Composition and Pseudo-Amino Acid Composition

Kuo-Chen Chou[1,2]* and Yu-Dong Cai[3,4]

[1]Gordon Life Science Institute, Torrey Del Mar Drive, San Diego, California 92130
[2]Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China
[3]Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai 200233, China
[4]Biomolecular Science Department, UMIST, P.O. Box 88, Manchester, M60IQD, UK

Abstract Recent advances in large-scale genome sequencing have led to the rapid accumulation of amino acid sequences of proteins whose functions are unknown. Since the functions of these proteins are closely correlated with their subcellular localizations, many efforts have been made to develop a variety of methods for predicting protein subcellular location. In this study, based on the strategy by hybridizing the functional domain composition and the pseudo-amino acid composition (Cai and Chou [2003]: Biochem. Biophys. Res. Commun. 305:407–411), the Intimate Sorting Algorithm (ISort predictor) was developed for predicting the protein subcellular location. As a showcase, the same plant and non-plant protein datasets as investigated by the previous investigators were used for demonstration. The overall success rate by the jackknife test for the plant protein dataset was 85.4%, and that for the non-plant protein dataset 91.9%. These are so far the highest success rates achieved for the two datasets by following a rigorous cross validation test procedure, further confirming that such a hybrid approach may become a very useful high-throughput tool in the area of bioinformatics, proteomics, as well as molecular cell biology. J. Cell. Biochem. 91: 1197–1203, 2004.

Key words: Intimate Sorting Algorithm; protein subcellular location; functional domain composition; pseudo-amino acid composition; InterPro database

Given the sequence of a protein, how can we predict which subcellular location it belongs to? This is currently a very hot topic in molecular and cell biology because the localization of a protein in a cell is closely correlated with its biological function (see, e.g., [Watson et al., 1987; Alberts et al., 1994; Lodish et al., 1995; Chou et al., 1998, 1999]). Also, the number of protein sequences entering into databanks has been rapidly increasing. It is anticipated that many more new protein sequences will be derived soon owing to the recent success of the human genome project, which has provided an enormous amount of genomic information in the form of 3 billion bp, assembled into tens of thousands of genes. In view of this, the importance to deal with such a problem is not only self-evident, but the challenge will also become even more critical and urgent in the near future. Actually, many efforts have been made trying to develop different computational methods for fast predicting the subcellular locations of proteins [Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994; Cedano et al., 1997; Claros et al., 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1999; Emanuellson et al., 2000; Pan et al., 2003; Zhou and Doctor, 2003]. Of these methods, some [Nakai and Kanehisa, 1992; Claros et al., 1997] were based on the N-terminal sorting signals. Their merit is with a clear biological implication because newly-synthesized proteins in vivo are governed by an intrinsic signal sequence to their destination, whether they are to pass through a membrane into a particular organelle, to become integrated into the membrane, or to be exported out of the cell [Blobel, 1976]. However, as pointed out by Reinhardt and Hubbard [1998], "in large

*Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130. E-mail: lifescience@san.rr.com

Received 12 October 2003; Accepted 24 October 2003
DOI 10.1002/jcb.10790

genome analysis projects genes are usually automatically assigned and these assignments are often unreliable for the 5′-regions." "This can lead to leader sequences being missing or only partially included, thereby causing problems for prediction algorithms depending on them." Therefore, most of the existing algorithms were actually based on the assumption that a protein could be represented by its amino acid composition derived from the entire sequence. However, the amino acid composition consists of only 20 components, each representing the occurrence frequency of 1 of the 20 native amino acids in a given protein. Thus, in the prediction algorithms developed from such an assumption, a protein would be formulated by a 20D (dimensional) vector [Chou and Zhang, 1993; Chou, 1995]. By so doing, all the sequence-order and sequence-length effects of a protein would be totally ignored and the prediction method based on the amino acid composition alone must bear a considerable intrinsic limitation. To avoid completely ignoring the contribution of the sequence-order effects, two different approaches, the so-called pseudo-amino acid composition approach [Chou, 2001] and the functional domain composition approach [Chou and Cai, 2002], was proposed.

The pseudo-amino acid composition consists of $20 + \lambda$ components, of which the first 20 components are the same as those in the conventional amino acid composition, and the components from $20 + 1$ to $20 + \lambda$ represent λ sequence-order correlation factors of different ranks. It is the additional λ components that have incorporated some sequence-order effects [Chou, 2001]. However, the pseudo-amino acid composition only includes the partial (but not complete) sequence-order information, and hence may still miss some information that might be immediately related to the function of a protein.

Subsequently, a completely different approach, the so-called functional domain composition [Chou and Cai, 2002] was proposed that contained the information of various functional domain types. The introduction of the functional domain composition represents an important progress in directly relating the localization of proteins with their function. Unfortunately, the current functional domain database [Murvai et al., 2001] is far from complete yet. Hence not all proteins can be properly defined by the database, leading to some setback in practical application [Chou and Cai, 2002].

In view of this, a strategy was proposed to represent a protein by hybridizing the functional domain composition and pseudo-amino acid composition [Cai and Chou, 2003]. The hybridization makes allowance for bringing out the best in each other. With such a strategy, the Intimate Sorting Algorithm (ISort predictor) is developed to predict the subcellular localization for plant and non-plant proteins, respectively, and the prediction quality has been further improved.

Hybridization of Functional Domain Composition and Pseudo-Amino Acid Composition

The original concept of the functional domain composition and the detailed procedure of how to use it to represent a protein were given in a previous study [Chou and Cai, 2002], where the functional domain composition was defined in the SBASE-A database [Murvai et al., 2001]. The SBASE-A database consists of 2,005 functional domains. With each of these domains as a base, a protein was defined as a 2005D vector in terms of its functional domain composition. In this study, the InterPro database, i.e., the integrated domain and motif database [Apweiler et al., 2001], was used to define the functional domain composition of a protein. InterPro release 5.2 (September 2002) contains 5,875 entries. With each of the 5,875 functional domains as a base, a protein can be defined as a 5875D vector, as illustrated by the following procedures.

(1) Use the program IPRSCAN [Apweiler et al., 2001] to search InterPro database for a given protein, if there is a hit (e.g., IPR001938, meaning the protein contains a sequence very similar to that of the 1938th domain of the InterPro database), then the 1938th component of the protein in the 5875D functional domain space is assigned 1; otherwise, 0.
(2) The protein can thus be explicitly formulated as

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_{5875} \end{bmatrix}, \qquad (1)$$

where

$$x_i = \begin{cases} 1, & \text{hit} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

It can be seen from the above equations that, instead of the 20D space [Chou and Zhang, 1993; Chou, 1995] in terms of the conventional amino acid composition, or the $(20+\lambda)$D space of the pseudo-amino acid composition [Chou, 2001], or the 2005D space of the functional domain composition [Chou and Cai, 2002] based on the SBASE-A database [Murvai et al., 2001], a protein is now defined in a 5875D space based on the InterPro database [Apweiler et al., 2001].

As mentioned above, since the current functional domain database is still far from complete, many proteins might not get any hits by following the above procedure and hence have no definition. For those proteins which could not be defined in the functional domain space, the pseudo-amino acid composition [Chou, 2001] was adopted to represent them. Using the pseudo-amino acid composition as a complemental approach has the following advantages: (1) a protein with a given sequence can always be uniquely defined; (2) some sequence-order effects can be taken into account. According to the concept of the pseudo-amino acid composition, a protein is formulated as [Chou, 2001]

$$\mathbf{X} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix}, \quad (3)$$

where the first 20 components are the same as those in the conventional amino acid composition and the components $p_{20+1}, \ldots, p_{20+\lambda}$ are associated with λ different ranks (Fig. 1) of sequence-order correlation factors as formulated by λ sub-equations of the following equation:

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} J_{i,i+1} \\ \tau_2 = \frac{1}{L-2}\sum_{i=1}^{L-2} J_{i,i+2} \\ \tau_3 = \frac{1}{L-3}\sum_{i=1}^{L-3} J_{i,i+3}, \quad (\lambda < L). \\ \cdots\cdots\cdots\cdots\cdots \\ \tau_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} J_{i,i+\lambda} \end{cases} \quad (4)$$

In the above equation, L is the chain length of the protein concerned, τ_1 is called the 1st-rank

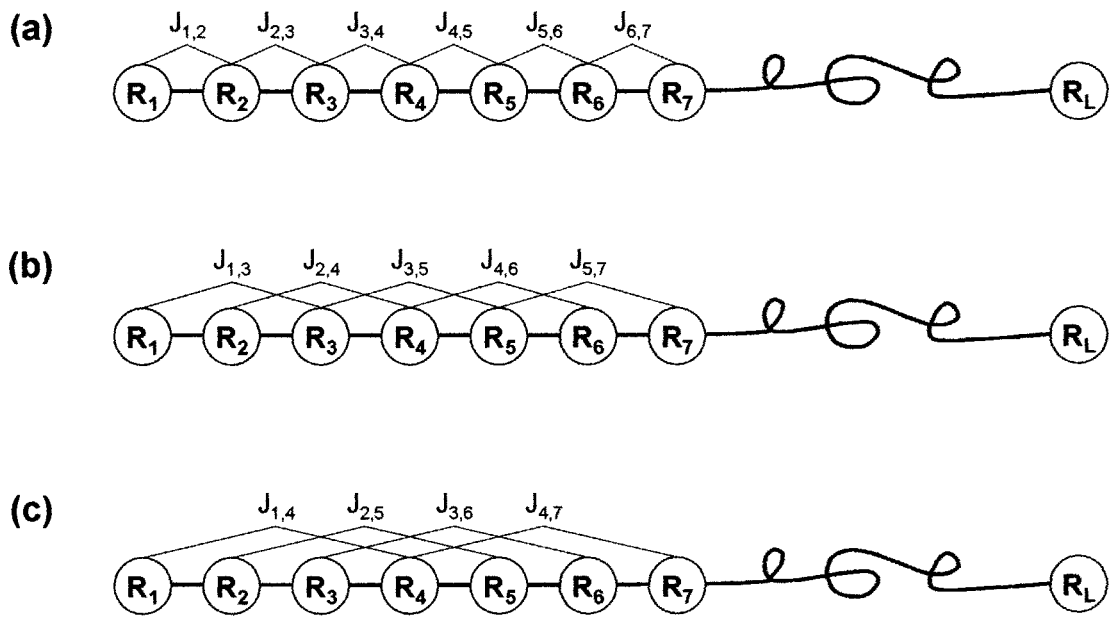

Fig. 1. A schematic drawing to show: (**a**) the 1st-rank, (**b**) the 2nd-rank, and (**c**) the 3rd-rank sequence-order correlation mode along a protein sequence. **Panel a**: Reflects the correlation mode between all the most contiguous residues; **panel b**: That between all the 2nd most contiguous residues; and **panel c**: That between all the 3rd most contiguous residues.

coupling factor that incorporates the sequence-order correlation between all the most contiguous residues along a protein chain (Fig. 1a), τ_2 the 2nd-rank coupling factor for all the 2nd most contiguous residues (Fig. 1b), τ_3 the 3rd-rank coupling factor for all the 3rd most contiguous residues (Fig. 1c), and so forth. The coupling factor $J_{i,j}$ in Equation 4 is a function of amino acids R_i and R_j that is defined by their characteristic quantities, such as the hydrophobicity value, hydrophilicity value, and side chain mass [Chou, 2001, 2002]. It can be seen from Figure 1 that the sequence-order effect of a protein is, to some degree, reflected via a set of discrete numbers $\tau_1, \tau_2, \tau_3, \ldots, \tau_\lambda$, as formulated in Equation 4. Actually, the first 20 components of Equation 1 reflect the conventional amino acid composition effect, while the components from $20+1$ to $20+\lambda$ reflect some sequence-order effect. A set of the $20+\lambda$ components as formulated by Equations 3–4 is called the pseudo-amino acid composition for protein X. Using such a name is because it still has the main feature of the conventional amino acid composition; but on the other hand, it contains the information beyond the conventional amino acid composition. Generally speaking, the larger the number of these correlation factors, the more the sequence-order effects incorporated. However, the number λ cannot exceed the length of a protein (i.e., the number of its total residues). For example, if a protein consists of 50 amino acid residues, then its pseudo-amino acid composition can contain as large as $20+50=70$ components, corresponding to a 70D vector. On the other hand, if the number of λ is too large, the overall success rate by jackknife tests might be reduced [Chou, 2001]. Therefore, for different training datasets, λ may have different optimal values. For the current study, the optimal value for λ is 27; i.e., the dimension of the pseudo-amino acid composition considered here is $20+27=47$. Given a protein, the 47 pseudo-amino acid components in Equation 3 can be easily derived by following the procedures as elaborated in the original study [Chou, 2001] that has first introduced the concept of pseudo-amino acid composition.

ISort Predictor

Suppose there are N proteins $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$ which have been classified into categories 1, $2, \ldots, \mu$. Now, for a query protein $\mathbf{X}$, how can we predict which category it belongs to? To deal with this problem, below we shall introduce a new algorithm, the so-called ISort predictor. First, let us define the similarity between $\mathbf{X}$ and $\mathbf{X}_i$ $(i=1,2,\ldots,N)$ given by

$$\Phi(\mathbf{X},\mathbf{X}_i)=\frac{\mathbf{X}\cdot\mathbf{X}_i}{\|\mathbf{X}\|\|\mathbf{X}_i\|},(i=1,2,\ldots,N) \quad (5)$$

where $\mathbf{X}\cdot\mathbf{X}_i$ is the dot product of vectors $\mathbf{X}$ and $\mathbf{X}_i$, and $\|\mathbf{X}\|$and$\|\mathbf{X}_i\|$ their modulus, respectively. Obviously, when $\mathbf{X}\equiv\mathbf{X}_i$, we have $\Phi(\mathbf{X},\mathbf{X}_i)=1$, meaning they have perfect or 100% similarity. Generally speaking, the similarity is within the range of 0 and 1; i.e., $0\le\Phi(\mathrm{X},\mathrm{X}_i)\le 1$. Accordingly, the ISort predictor can be formulated as follows. If the similarity between $\mathbf{X}$ and $\mathbf{X}_k(k=1,2,\ldots,$ or $N)$ is the highest; i.e., Equation 6

$$\Phi(\mathrm{X},\mathrm{X}_k) = \mathrm{Max}\{\Phi(\mathrm{X},\mathrm{X}_1),\Phi(\mathrm{X},\mathrm{X}_2),\ldots,\Phi(\mathrm{X},\mathrm{X}_N)\} \quad (6)$$

then the query protein $\mathbf{X}$ is predicted belonging to the same category as of $\mathbf{X}_k$. If there is a tie, the query protein is not uniquely determined, but cases like that rarely occur.

As mentioned above, since the query protein may or may not get a hit in search the InterPro database [Apweiler et al., 2001]. If the query protein has no hit found during the prediction process (cf. Equation 2), it cannot be defined (except for a null vector) in the 5875D functional domain composition space (cf. Equation 1). Under such a circumstance, the query protein, as well as all the proteins in the training dataset, should be defined in the $(20+\lambda)\mathrm{D}=47\mathrm{D}$ pseudo-amino acid composition space as given by Equation 3 in operating the ISort predictor. However, if the query protein can be defined in the 5875D functional domain composition as given by Equation 1, the prediction should be carried out based on all those proteins in the training set that can be defined in the same 5875D space as well. Accordingly, the current ISort predictor actually consists of two sub predictors: (1) the ISort-5875D predictor that operates in the 5875D functional domain composition space, and (2) the ISort-47D predictor that operates in the 47D pseudo-amino acid composition space with $\lambda=27$.

RESULTS AND DISCUSSION

To benchmark the prediction quality of the current method against others and make the comparison more objectively, the datasets

constructed by other investigators were used for demonstration. The datasets constructed by Emanuellson et al. [2000] contain two redundancy-reduced sets. One is the plant set that consists of 940 proteins, of which 141 are destined for the chloroplast, 368 for the mitochondrion, 269 for the secretory pathway, and 162 for the other localizations such as nuclear and cytosolic. The other is the non-plant set that consists of 2,738 proteins, of which 371 are destined for the mitochondrion, 715 for the secretory pathway, and 1,652 for the other localizations such as nuclear and cytosolic. According to the report by these authors, the overall success rate predicted by the TargetP predictor [Emanuellson et al., 2000] for the 940 plant proteins classified into four different categories was 85%, and that for the 2,738 non-plant proteins classified into three categories was 90%. These are the highest success rates so far reported for the aforementioned plant and non-plant datasets. Now for exactly the same datasets, we used ISort predictor to perform prediction.

The computation was carried out in a Silicon Graphics IRIS Indigo workstation (Elan 4000). By searching the InterPro database for the 940 protein sequences in the plant set, 745 sequences got hits, and 195 did not. And for the 2,738 protein sequences in the non-plant set, 2,423 got but 315 not. This means that, if only the functional domain composition approach was used, 195 proteins in the plant set and 315 in the non-plant set would have no definition, leading to a failure of identifying their localization. That is why it is important to hybridize it with the pseudo-amino acid composition approach, by which not only a protein can always be defined but also its sequence-order effects may considerably be reflected. Thus, the hybrid algorithm was operated according to the following flowchart: if a query protein got a hit by search InterPro database, then the ISort-5875D predictor was used to predict its subcellular location; otherwise, the ISort-47D predictor was used for the prediction.

The prediction quality was examined by the jackknife test. For the convenience of readers, a brief introduction of jackknife test procedure is given below. During the process of jackknife tests for the current study, each protein in the plant or non-plant set is in turn singled out as a query protein and all the rule-parameters are computed based on the remaining proteins. In other words, the subcellular location of each query protein is identified by the rule parameters derived from all the other proteins except the query one. During the process of jackknifing both the training dataset and testing dataset are actually open, and a protein will in turn move from one to the other until all the proteins have been identified.

Compared with the independent dataset test and sub-sampling test which are often adopted in the literature of biology, the jackknife test is thought the most objective and effective method for cross-validation in statistical prediction [Zhou, 1998; Zhou and Assa-Munt, 2001]; see, e.g., a monograph [Mardia et al., 1979] for the mathematical principle and a comprehensive review [Chou and Zhang, 1995] in this regard. This is because in the independent dataset test, the selection of a testing dataset is quite arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing dataset is sufficiently large. As for the sub-sampling test in which a given dataset is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Emanuellson et al. [2000], proteins in each group were "truncated to a number divisible by five" and then "divided into five equally sized parts for cross-validation." Four of them were used as the training data and one as the testing data. Thus, the number of possible divisions for the plant set would be $\Pi = \Pi_1 \times \Pi_2 \times \Pi_3 \times \Pi_4$, where $\Pi_1 = \frac{140!}{28!28!28!28!28!5!}$, $\Pi_2 = \frac{365!}{73!73!73!73!73!5!}$, $\Pi_3 = \frac{265!}{53!53!53!53!53!5!}$, and $\Pi_4 = \frac{160!}{32!32!32!32!32!5!}$. Of $\Pi_1, \Pi_2, \Pi_3,$ and Π_4, the smallest is $\Pi_1 \approx 4 \times 10^{91}$, indicating the number of possible divisions would be $\Pi \gg 10^{365}$. This is an astronomical figure, which is too large to be handled by any existing computers. Hence in any practical sub-sampling tests as conducted by Emanuellson et al. [2000], only a very small fraction of the possible divisions were investigated, and the results thus obtained would be quite arbitrary and might be overestimated, as will be further discussed later. Accordingly, the jackknife test as adopted here is much more objective and rigorous. The overall success rates thus obtained are given in Table I. For facilitating comparison, the rates obtained by the other predictors, such as TargetP [Emanuellson et al., 2000], and Psort [Nakai and Kanehisa, 1992; Nakai and Horton, 1999], are also listed in the same table. From Table I we can see the

TABLE I. Comparison of Localization Predictor Performances on the Redundancy-Reduced 940 Plant Proteins and 2,738 Non-Plant Proteins

	Overall success rate			
	Plant		Non-plant	
Predictor	Jackknife (%)	Sub-sampling (%)	Jackknife (%)	Sub-sampling (%)
ISort[a]	85.4	92.3	91.9	98.3
TargetPb[b]	N/A	85.0	N/A	90.0
Psort[c]	N/A	69.8	N/A	83.2

[a]The present paper.
[b]TargetP is a neural network-based tool for protein subcellular location prediction using N-terminal sequence information only [Emanuellson et al., 2000].
[c]PSort is a program widely used for detecting sorting signals in proteins and predicting their subcellular localization [Nakai and Horton, 1999; Nakai and Kanehisa, 1992].

following: (1) the overall success rates obtained with ISort, which has combined both the functional domain and sequence-order effects, are significantly higher than those from the other predictors. For example, the overall success rates yielded from ISort predictor by the same sub-sampling test procedure as described in Emanuellson et al. [2000] are 92.3 and 98.3% for the plant set and non-plant set, respectively, which are more than 7 and 8% higher than the corresponding rates by TargetP, indicating that the subcellular localization of a protein is closely related to its function in both the plant and non-plant cases; (2) the success rates obtained with ISort by jackknife tests are remarkably lower than those by sub-sampling tests. This is fully consistent with the fact that the results obtained by sub-sampling tests could not avoid arbitrariness and might be overestimated, as mentioned above as well as in many previous publications (see, e.g., [Mardia et al., 1979; Chou and Zhang, 1995; Zhou, 1998; Zhou and Assa-Munt, 2001]). Even though, the overall success rates yielded from ISort predictor by jackknife tests are still higher than those from the other predictors by sub-sampling tests for both the plant and non-plant protein datasets, which is a compelling evidence to indicate the superiority of the present approach.

Accordingly, from both the rationality of testing procedure and the success rates of test results, the hybridization of the functional domain composition and pseudo-amino acid composition as presented in this study can significantly improve the prediction quality of subcellular localization of proteins.

CONCLUSION

The pseudo-amino acid composition approach [Chou, 2001] and the functional domain composition approach [Chou and Cai, 2002] are two completely different approaches developed for improving the prediction quality of protein subcellular location. They are both quite powerful, but each has its own limitation. The present study has demonstrated that a hybridization of the two different approaches can make them complement each other, and that the introduction of the ISort predictor can make allowance for bringing out the best in each other and making each shining more brilliantly in the other's company. This is the essence why the current method is superior to others in predicting subcellular localization of proteins. It is instructive to point out that the current approach can also be used to improve the prediction quality for other protein attributes [Chou, 2002], such as enzyme family classes [Chou and Elrod, 2003] and protein quaternary structure attributes [Chou and Cai, 2003].

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this article.

REFERENCES

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. 1994. Molecular biology of the cell, chap. 1. New York & London: Garland Publishing.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy L, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A,

Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research 29:37–40.

Blobel G. 1976. Extraction from free ribosomes of a factor mediating ribosome detachment from rough microsomes. Biochem Biophys Res Comm 68:1–7.

Cai YD, Chou KC. 2003. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm 305: 407–411.

Cedano J, Aloy P, P'erez-Pons JA, Querol E. 1997. Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600.

Chou JJ, Li H, Salvesen GS, Yuan J, Wagner G. 1999. Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 96:615–624.

Chou JJ, Matsuo H, Duan H, Wagner G. 1998. Solution structure of the RAIDD CARD and model for CARD/ CARD interaction in caspase-2 and caspase-9 recruitment. Cell 94:171–180.

Chou JJ, Zhang CT. 1993. A joint prediction of the folding types of 1,490 human proteins from their genetic codons. J Theor Biol 161:251–262.

Chou KC. 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Struct Funct Genet 21:319–344.

Chou KC. 2001. Prediction of protein cellular attributes using pseudo-amino-acid-composition, (Erratum Vol. 44, 60). Proteins: Struct Funct Genet 43:246–255.

Chou KC. 2002. A new branch of proteomics: Prediction of protein cellular attributes. In: Weinrer PW, Lu Q, editors. Gene cloning & expression technologies. Westborough, MA: Eaton Publishing. pp 57–70.

Chou KC, Cai YD. 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769.

Chou KC, Cai YD. 2003. Predicting protein quaternary structure by pseudo amino acid composition. Proteins 53:282–289.

Chou KC, Elrod DW. 1999. Protein subcellular location prediction. Protein Engineering 12:107–118.

Chou KC, Elrod DW. 2003. Prediction of enzyme family classes. J Proteome Res 2:183–190.

Chou KC, Zhang CT. 1995. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349.

Claros MG, Brunak S, von Heijne G. 1997. Prediction of N-terminal protein sorting signals. Curr Opin Struct Biol 7:394–398.

Emanuellson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016.

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. 1995. Molecular cell biology, chap. 3. New York: Scientific American Books.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. London: Academic Press. pp 322–381.

Murvai J, Vlahovicek K, Barta E, Pongor S. 2001. The SBASE protein domain library, release 8.0: A collection of annotated protein sequence segments. Nucleic Acids Research 29:58–60.

Nakai K, Horton P. 1999. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–36.

Nakai K, Kanehisa M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14:897–911.

Nakashima H, Nishikawa K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238:54–61.

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L. 2003. Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. J Protein Chem 22:395–402.

Reinhardt A, Hubbard T. 1998. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26:2230–2236.

Watson JD, Hopkins NH, Roberts JW, Steitz JA, Weiner AM. 1987. Molecular biology of the gene 4th edn. Menlo Park, CA: Benjamin/Cummings Publishing Company, Inc.

Zhou GP. 1998. An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738.

Zhou GP, Assa-Munt N. 2001. Some insights into protein structural class prediction. Proteins: Struct Funct Genet 44:57–59.

Zhou GP, Doctor K. 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct Funct Genet 50: 44–48.